Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis

Gangu DharmaRaju [1], Srikanth Murapaka [2], M. JayanthiRao [3]

1Final Year M. Tech-CSE, Sarada Institute of Science Technology and Management, Ampolu, Srikakulam, India.
2Asst. Prof., Dept of CSE, Sarada Institute of Science Technology and Management, Ampolu, SKLM, India.
3Associate. Prof. and Head of the Dept. of CSE, Sarada Institute of Science Technology and Management, SKLM, India.

ABSTRACT

Data mining is widely used domain for extracting trends or patterns from historical data. However, the databases used by enterprises can’t be directly used for data mining. It does mean that Data sets are to be prepared from real world database to make them suitable for particular data mining operations. However, preparing datasets for analyzing data is tedious task as it involves many aggregating columns, complex joins, and SQL queries with sub queries. More over the existing aggregations performed through SQL functions such as MIN, MAX, COUNT, SUM, AVG return a single value output which is not suitable for making datasets meant for data mining. In fact these aggregate functions are generating vertical aggregations. This paper presents techniques to support horizontal aggregations through SQL queries. The result of the queries is the data which is suitable for data mining operations. It does mean that this paper achieves horizontal aggregations through some constructs built that includes SQL queries as well. The methods prepared by this paper include CASE, SPJ and PIVOT. We have developed a prototype application and the empirical results reveal that these constructs are capable of generating data sets that can be used for further data mining operations.

Index Terms – Aggregations, SQL, data mining, OLAP, and data set generations

1. INTRODUCTION

RDBMS has become a de facto standard for storing and retrieving large amount of data. This data is permanently stored and retrieved through front end applications. The applications can use SQL to interact with relational databases. Preparing databases needs identification of relevant data and then normalizing the tables. Aggregations are supported by SQL to obtain summary of data. The aggregate functions supported by SQL are SUM, MIN, MAX, COUNT and AVG. These functions produce single value output. These are known as vertical aggregations. This is because each function operates on the values of a domain vertically and produces a single value result. The result of vertical aggregations is useful in calculations or computations. However, they can’t be directly used in data mining operations further. In fact summary data sets can be prepared and they can be used further in data mining operations .The summary data can also be used in statistical algorithms. Most of the data mining operations expect a data set with horizontal layout with many tuples and one variable or dimension per column. This is the case with many data mining algorithms like PCA, regression, classification, and clustering

As horizontal aggregations are capable of producing data sets that can be used for real world data mining activities, this paper presents three horizontal aggregations namely SPJ, PIVOT and CASE. It does mean that we enhanced these operators that are provided by SQL in one way or the other. The SPJ aggregation is developed using construct with standard SQL operations. PIVOT makes use of built in pivoting facility in SQL while the CASE is built based on the SQL CASE construct. We have built a web based prototype application that demonstrates the effectiveness of these constructs. The empirical results revealed that these operations are able to produce data sets with horizontal layout that is suitable for OLAP operations or data mining operations. The motivation behind this work is that developing data sets for data mining operations is very difficult and time consuming. One problem in this area is that the existing SQL aggregations provide a single row output. This output is not suitable for data mining operations. For this reason we have extended the functionalities of CASE, SPJ and PIVOT operators in such a way that they produce data sets with horizontal layouts.

The proposed horizontal aggregations have some unique features and they are very useful. The advantages include, they provide construct that can generate SQL code that results in dataset which is suitable for data mining; the SQL code supports automation of writing SQL queries, testing them and optimizing them. As the proposed constructs are based on SQL, it reduces lot of coding as it is a powerful data retrieval language. The proposed system is user friendly as users are never expected to write queries. Instead, they just make queries in a user-friendly fashion. End users who do not know SQL also can make use of the proposed system. As SQL code is automatically generated based on the operator used in the query, it reduces lot of manual mistakes. In modern database where data is stored because of day to day operations, users
do not get a chance to use data directly for mining operations. Instead, it has to be transformed to make sense for data mining operations. Generally data from business database is converted, and loaded into some other data sets known as data warehouse. The proposed horizontal aggregations can be used to generate data sets for the purpose of data mining analysis.

2. HORIZONTAL AGGREGATIONS
For describing the methods pertaining to the proposed horizontal aggregations such as PIVOT, CASE and SPJ, input table as shown in Fig. 1 (a), traditional vertical sum aggregation as shown in Fig 1 (b) and horizontal aggregation as shown in Fig. 1 (c) are considered.

As can be seen in fig. 1, input table has some sample data. Traditional vertical sum aggregations are presented in (b) which is the result of SQL SUM function while (c) holds the horizontal aggregation which is the result of SUM function.

**STEPS USED IN ALL METHODS**

As can be seen in fig. 2, for all methods such as SPJ, CASE and PIVOT steps are given. For every method the procedure starts with SELECT query. Afterwards, corresponding operator through underlying construct is applied. Then horizontal aggregation is computed.

As can be seen in fig. 3, for all methods such as SPJ, CASE and PIVOT steps are given. For every method the procedure starts with SELECT query. Afterwards, Corresponding operator through underlying construct is applied. Then horizontal aggregation is computed.
• **SPJ Method**

This method is based on the relational operators only. In this method one table is created with vertical aggregation for each column. Then all such tables are joined in order to generate a table containing horizontal aggregations. The actual implementation is based on the details given in.

• **PIVOT Method**

This aggregation is based on the PIVOT operator available in RDBMS. As it can provide transpositions, it can be used to evaluate horizontal aggregations. PIVOT operator determines how many columns are required to hold transpose and it can be combined with GROUP BY clause.

```sql
Listing 1 – Shows optimized instructions for PIVOT construct

As can be seen in listing 1, the optimized query projects only the columns that participate in computation of horizontal aggregations.

• **CASE Method**

This construct is built based on the existing CASE structure of SQL. Based on Boolean expression one of the results is returned by CASE construct. It is same as projection/aggregation query from relational point of view. Based on some injunction of conditions each non key value is given by a function. Here two basic strategies to compute horizontal aggregations. The first strategy is to compute directly from input table. The second approach is to compute vertical aggregation and save the results into temporary table. Then that table is further used to compute horizontal aggregations.

3. **Experimental Evaluation**

The environment used for the development of prototype web based application that demonstrates the three horizontal aggregations include Visual Studio 2010, and SQL Server 2008. The former is used to build front end application with web based interface while the latter is used to store data permanently. The technologies used include ASP.NET which is meant for developing web services and web applications, and AJAX (Asynchronous JavaScript and XML) for rich user experience. Programming language used in C# which is an object oriented high level programming language. The result of horizontal aggregation SPJ is shown in fig.

![Fig. 1 – Results of SJP aggregation](image-url)
As can be seen in fig. 1, the results are through the SPJ operation that results in data in horizontal layout. Data in this layout can be considered as data set that can be used for further data mining operations.

![Fig. 2 – Result of Pivoting Aggregation](image)

As can be seen in fig. 2, the results are through the PIVOT operation that results in data in horizontal layout. Data in this layout can be considered as data set that can be used for further data mining operations.

![Fig. 3 – Result of CASE Aggregation](image)

As can be seen in fig. 2, the results are through the CASE operation that results in data in horizontal layout. Data in this layout can be considered as data set that can be used for further data mining operations.

4. CONCLUSIONS
In this paper we extended three aggregate functions such as CASE, SPJ and PIVOT. These are known as horizontal aggregations. We have achieved it by writing underlying constructs for each operator. When they are used, internally the corresponding construct gets executed and the resultant data set is meant for OLAP (Online Analytical Processing). Vertical aggregations such as SUM, MIN, MAX, COUNT, and AVG return a single value output. However, that output can’t be used for data mining operations. In order to prepare real World datasets that are very much suitable for data mining operations, we explored horizontal aggregations by developing constructs in the form of operators such as CASE, SPJ and PIVOT. Instead of single value, the horizontal aggregations return a set of values in the form of a row. The result resembles a multidimensional vector. We have implemented SPJ using standard relational query operations. The CASE construct is developed extending SQL CASE. The PIVOT makes use of built in operator provided by RDBMS for pivoting data. To evaluate these operators, we have developed a web based prototype application and results reveal that the proposed horizontal aggregations are capable of preparing data sets for real world data mining operations.

REFERENCES


[18] Carlos Ordonez and Zhibo Chen, “Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis”, IEEE

Author

Gangu DharmaRaju is student of Sarada Institute of Science Technology and Management, Ampolu Rd., Srikakulam, AP, India. He has received B.Tech information technology M.Tech Degree in CSE. His main research interest includes data mining.

Srikanth Murapaka is working as Assistant Prof. at Sarada Institute of Science Technology and Management, Ampolu Rd., Srikakulam, AP, India. He has received M.Tech Degree in CSE. His Main Research interest includes Data Mining.

Jayanthi Rao Madina is working as a HOD in Sarada Institute of Science, Technology And Management, Srikakulam, Andhra Pradesh. He received his M.Tech Degree in CSE. His Main Research interest includes Image Processing, Computer Networks, Data Mining, Distributed Systems.